

# “Compressed” Compressed Sensing

Galen Reeves and Michael Gastpar

Department of Electrical Engineering and Computer Sciences

University of California, Berkeley

Berkeley, CA, 94720, USA

Email: {greeves, gastpar}@eecs.berkeley.edu

**Abstract**—The field of compressed sensing has shown that a sparse but otherwise arbitrary vector can be recovered exactly from a small number of randomly constructed linear projections (or samples). The question addressed in this paper is whether an even smaller number of samples is sufficient when there exists prior knowledge about the distribution of the unknown vector, or when only partial recovery is needed. An information-theoretic lower bound with connections to free probability theory and an upper bound corresponding to a computationally simple thresholding estimator are derived. It is shown that in certain cases (e.g. discrete valued vectors or large distortions) the number of samples can be decreased. Interestingly though, it is also shown that in many cases no reduction is possible.

## I. INTRODUCTION

Suppose that an unknown vector  $\mathbf{x}$  of length  $n$  is observed using a set of linear projections  $\mathbf{y} = A\mathbf{x}$  where  $A$  is a known  $m \times n$  sampling matrix. The field of compressed sensing (see references in [1]) has shown that if  $\mathbf{x}$  is sparse (i.e. has a relatively small number of nonzero elements) then exact recovery is possible even if the number of samples  $m$  is much less than the vector length  $n$ . A great deal of work has considered necessary and sufficient conditions on the sampling matrix  $A$  with respect to various recovery goals. In particular, much of this work has focused on sufficient conditions for computationally efficient recovery algorithms.

Typically, the conditions on the sampling matrix are remarkably general with respect to the unknown vector  $\mathbf{x}$  in the sense that they require no assumptions about the values or locations of the nonzero elements. Moreover, many of the results still apply even if  $\mathbf{x}$  is not actually sparse, but instead has a sparse representation with respect to a known basis.

In many practical situations however, there exists prior knowledge about the values of the nonzero elements. In this paper, we address the extent to which this additional information allows for recovery using an even smaller number of samples than are needed in the general “compressed sensing” setting. We focus exclusively on recovery of the support set (i.e. the locations of the nonzero elements) in the high dimensional setting and ask the following two questions:

- *What if we consider approximate support recovery?*

In Section III we show that if the sampling matrix  $A$  is designed with knowledge of the basis in which  $\mathbf{x}$  is sparse, then there exists a natural tradeoff between accuracy and the number of samples. Conversely, if the sampling matrix is designed independently of the sparse basis, then no such tradeoff is possible.

- *What if  $\mathbf{x}$  is a random vector with a known distribution?*

If the distribution is discrete, then it is straightforward to see that only one sample is needed. In Section IV we consider general distributions, and our main results (Theorems 1 and 2) show that knowledge of the distribution may or may not decrease the number of samples that are needed depending on the desired distortion and various properties of the distribution such as the differential entropy.

An additional contribution of this paper is given by the proof of our main lower bound (Section V) which uses results from free probability theory to characterize the limiting distributions of certain random matrices that occur frequently in compressed sensing.

A number of related works have addressed various bounds on the asymptotic sampling rate needed for the noisy setting [2]–[12]. In these cases, it is clear that properties such as the size of the smallest nonzero values dramatically affect the number of samples that are needed. The noiseless setting addressed in the paper however, gives insight about fundamental limitations of the sampling process that cannot be overcome simply by increasing the signal to noise ratio.

## II. PROBLEM SETUP

We consider a generalized sparsity model where an unknown vector  $\mathbf{x} \in \mathbb{R}^n$  is assumed to have a sparse representation  $\mathbf{u} \in \mathbb{R}^n$  with respect to a known orthonormal basis  $B \in \mathbb{R}^{n \times n}$  given by

$$\mathbf{x} = B\mathbf{u}.$$

The *support*  $\mathbf{s} \subset \{1, 2, \dots, n\}$  is the set of integers indexing the nonzero elements of  $\mathbf{u}$ ,

$$\mathbf{s} := \{i : u_i \neq 0\},$$

and the sparsity  $k = |\mathbf{s}|$  is the number of nonzero elements.

The vector of samples  $\mathbf{y} \in \mathbb{R}^m$  is expressed in terms of a sampling matrix  $A \in \mathbb{R}^{m \times n}$ :

$$\mathbf{y} = A\mathbf{x}.$$

Throughout this paper, we assume that an estimator is given the set  $(\mathbf{y}, A, B, k)$  and the goal is to recover the support  $\mathbf{s}$  of the sparse representation  $\mathbf{u}$ . The distortion between a support  $\mathbf{s}$  and its estimate  $\hat{\mathbf{s}}$  is measured using the Hamming distance

$$d(\mathbf{s}, \hat{\mathbf{s}}) := |\mathbf{s} \cup \hat{\mathbf{s}}| - |\mathbf{s} \cap \hat{\mathbf{s}}|.$$

This paper focuses on whether or not a given recovery task is possible using an  $m \times n$  sampling matrix  $A$ . One possible requirement is that  $A$  be good uniformly for all possible  $k$ -sparse vectors. However, this paper considers a less stringent requirement and instead asks if there exists a distribution  $p_A$  such that recovery is possible, with high probability, for any  $k$ -sparse vector  $\mathbf{u}$  when  $\mathbf{A} \sim p_A$  is a random matrix drawn independently of  $\mathbf{u}$ , and possibly also  $B$ .

To highlight the difference between the above requirements it is useful to consider the task of exact recovery. Then, it can be shown that there exists a sampling matrix  $A$  satisfying the first requirement if and only if  $m \geq \min(2k, n)$ , whereas there exists a distribution  $p_A$  satisfying the second requirement if and only if  $m \geq \min(k+1, n)$ .

To characterize the number of samples that are needed, we focus on the high dimensional setting where the vector length  $n$  becomes large. We assume that for each  $n$ , the sparsity is given by  $k_n = \lfloor \Omega \cdot n \rfloor$  for some known *sparsity rate*  $\Omega \in (0, 1)$ . The following definitions are used to characterize the asymptotic *sampling rate* given by  $\rho = m_n/n$ .

**Definition 1.** The general source  $\mathcal{X}^n(\Omega)$  outputs an arbitrary (non-random) vector  $\mathbf{x} \in \mathbb{R}^n$  and basis  $B \in \mathbb{R}^{n \times n}$  where  $\mathbf{x} = B\mathbf{u}$  for some vector  $\mathbf{u} \in \mathbb{R}^n$  whose support  $\mathbf{s}$  has size  $k = \lfloor \Omega \cdot n \rfloor$ .

Given any support estimator  $\hat{\mathbf{s}}(\mathbf{y}, A, B, k)$  and any distribution  $p_A$ , the probability that the fraction of errors exceeds the normalized distortion  $\alpha \in [0, 1]$  for the general source  $\mathcal{X}^n(\Omega)$  is given by

$$P_e^{(n)} = \inf_{(\mathbf{x}, B) \in \mathcal{X}^n(\Omega)} \Pr\left\{d(\mathbf{s}, \hat{\mathbf{s}}(\mathbf{y}, \mathbf{A}, B, k)) > \alpha \cdot k\right\}.$$

**Definition 2.** A sampling rate distortion pair  $(\rho, \alpha)$  is said to be *achievable* for a source  $\mathcal{X}$  if for each integer  $n$  there exists an estimator  $\hat{\mathbf{s}}(\mathbf{y}, A, B, k)$  and a distribution  $p_A$  on a  $\lceil \rho \cdot n \rceil \times n$  sampling matrix such that

$$P_e^{(n)} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

The *sampling rate distortion function*  $\rho(\alpha)$  is the infimum of rates  $\rho \geq 0$  such that the pair  $(\rho, \alpha)$  is achievable.

### III. ARBITRARY SIGNALS

This section considers the sampling rate distortion function  $\rho(\alpha)$  of the general source  $\mathcal{X}(\Omega, F)$  for two different restrictions on the sampling matrix.

**Definition 3.** A random sampling matrix  $\mathbf{A}$  is said to be *universal* if it is drawn independently of the basis  $B$ , and *basis-specific* otherwise.

One useful property of a universal sampling matrix is that the sampling matrix can be constructed without knowledge of the sparse basis. Recovery with respect to a basis-specific matrix, however, is equivalent to assuming the the basis is the identity matrix (i.e.  $B = I$ ) since any target matrix  $\mathbf{A}_0$  designed for this setting can be applied to a general basis  $B$  by using the sampling matrix  $\mathbf{A} = \mathbf{A}_0 B^{-1}$ . The following

result shows that the universal and basis-specific settings are the same when exact recovery is required but significantly different when a nonzero distortion is allowed.

**Proposition 1.** The sampling rate distortion function  $\rho(\alpha)$  of the general source  $\mathcal{X}(\Omega)$  is given by

$$\rho(\alpha) = \begin{cases} \Omega, & \text{if } \mathbf{A} \sim p_A \text{ is universal} \\ (1 - \alpha)\Omega, & \text{if } \mathbf{A} \sim p_{A|B} \text{ is basis-specific} \end{cases} \quad (1)$$

for  $\alpha < 1$  and is equal to zero otherwise.

*Proof Sketch:* If the basis is known, then a “rate sharing” strategy may be employed to convexify the achievable rate distortion region. Roughly speaking, this corresponds to ignoring some randomly chosen subset of the elements of  $\mathbf{u}$  by placing zeros in the corresponding columns of the matrix  $AB$ . In the universal setting, however, this strategy is not possible. A full proof is given in [11]. ■

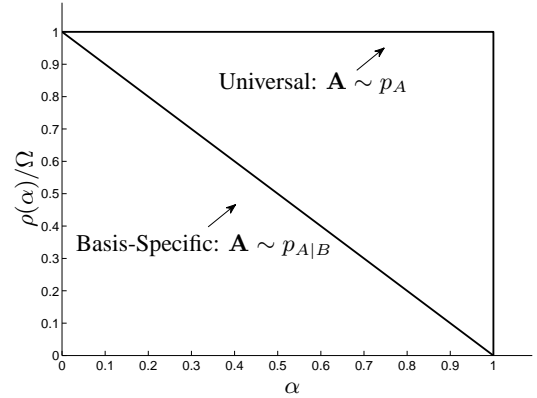


Fig. 1. Comparison of the normalized sampling rate distortion function  $\rho(\alpha)/\Omega$  of the general source  $\mathcal{X}(\Omega)$  as a function of the distortion  $\alpha$  for the universal and basis-specific settings.

### IV. RANDOM SIGNALS

So far, we have considered the recovery of arbitrary vectors and the results have been mostly algebraic. In this section, we consider recovery of random vectors. We focus exclusively on the universal setting where the sampling matrix  $\mathbf{A}$  must be designed independently of the sparse basis  $B$ .

**Definition 4.** The random source  $\mathcal{X}^n(\Omega, F)$  outputs a random vector  $\mathbf{X} \in \mathbb{R}^n$  and basis  $\mathbf{B} \in \mathbb{R}^{n \times n}$  where  $\mathbf{X} = \mathbf{B}\mathbf{U}$  for a random vector  $\mathbf{U} \in \mathbb{R}^n$  whose support  $\mathbf{S}$  is distributed uniformly over all possibilities of size  $k = \lfloor \Omega \cdot n \rfloor$  and whose nonzero elements  $\{U_i : i \in \mathbf{S}\}$  are i.i.d.  $\sim F$ . The basis  $\mathbf{B}$  is distributed uniformly over the set of all orthonormal matrices and is independent of  $\mathbf{U}$ .

We assume throughout that  $F$  denotes the distribution of a real valued random variable with finite power and zero probably mass at zero. Also, the definitions of achievability are the same as for the general source, except that the probability of error is taken with respect to the random vector  $\mathbf{X}$  and random basis  $\mathbf{B}$ ,

$$P_e^{(n)} = \Pr\left\{d(\mathbf{S}, \hat{\mathbf{s}}(\mathbf{Y}, \mathbf{A}, \mathbf{B}, k)) > \alpha \cdot k\right\}. \quad (2)$$

### A. Lower Bounds

This section gives an information theoretic lower bound on the sampling rate distortion function  $\rho(\alpha)$  of a random source  $\mathcal{X}(\Omega, F)$ . To begin, we note that in some cases, the constraints imposed by the distribution  $F$  significantly alter the nature of the estimation problem.

**Proposition 2** (Discrete Signals). *Suppose that the distribution  $F$  is supported on a discrete and finite set  $\Sigma \subset \mathbb{R} \setminus \{0\}$ . Then, only  $m = 1$  sample is sufficient for exact recovery, and the sampling rate distortion function  $\rho(\alpha)$  of the random source  $\mathcal{X}(\Omega, F)$  is  $\rho(\alpha) = 0$  for all  $\alpha$ .*

*Proof:* Suppose that  $\mathbf{A}$  is an  $1 \times n$  “matrix” whose elements are drawn i.i.d. from continuous distribution with finite power. Then, with probability one, the projection  $\mathbf{u} \mapsto \mathbf{A}\mathbf{B}\mathbf{u}$  maps each of the  $\binom{n}{k} k^{|\Sigma|}$  possible realizations of  $\mathbf{u}$  to a unique real number. ■

The fact that only one sample is needed for discrete distributions is not due to the sparsity in the problem (after all, the result does not depend on the sparsity rate  $\Omega$ ) and Proposition 2 provides little insight into cases where the unknown signal may have a density. To address these cases, we introduce the following property of a random signal source.

**Definition 5.** Given any distribution  $F$  with a density and any sparsity rate  $\Omega$  the function  $\theta(\Omega, F) \in [0, 1]$  is given by

$$\theta(\Omega, F) = \frac{(2\pi e)^{-1} \exp(2h(F))}{\sigma_F^2 + (1 - \Omega) \mu_F^2}, \quad (3)$$

where  $\mu_F$ ,  $\sigma_F^2$ , and  $h(F)$  denote the mean, variance and differential entropy of the distribution  $F$ . If  $F$  does not have a density, then  $\theta(\Omega, F) \equiv 0$ .

The property  $\theta(\Omega, F)$  is the normalized entropy power of the nonzero elements and is equal to one if and only if  $F$  is a zero mean Gaussian distribution. Roughly speaking, one may interpret  $\theta(\Omega, F)$  as the relative “distance” between a random source  $\mathcal{X}(\Omega, F)$  and a discrete source. The following result, which is proved in Section V, uses this property to lower bound the sampling rate distortion function.

**Theorem 1** (Lower Bound). *A sampling rate distortion pair  $(\rho, \alpha)$  is not achievable for the random source  $\mathcal{X}(\Omega, F)$  if  $\rho < \Omega$  and*

$$\frac{\rho}{2} \log \left( \frac{1}{\theta(\Omega, F)} \cdot \frac{\Delta(\rho)}{\Delta(\rho/\Omega)} \right) < H(\Omega) - H(\alpha\Omega) \quad (4)$$

where  $\theta(\Omega, F)$  is given by Definition 5,  $H(p) = -p \log(p) - (1-p) \log(1-p)$  is binary entropy and

$$\Delta(r) = \begin{cases} (1-r)^{1-1/r} & \text{if } r < 1 \\ 1 & \text{if } r = 1 \end{cases}. \quad (5)$$

One consequence of Theorem 1, is that there is a simple test to see whether or not the sampling rate needed for a random source  $\mathcal{X}(\Omega, F)$  is any less than that needed for the general source  $\mathcal{X}(\Omega)$ .

**Corollary 1** (Theorem 1). *The sampling rate distortion function  $\rho(\alpha)$  of the random source  $\mathcal{X}(\Omega, F)$  is given by  $\rho(\alpha) = \Omega$  for all  $\alpha < 1$  such that*

$$\theta(\Omega, F) > \Delta(\Omega) \exp \left( -\frac{2}{\Omega} [H(\Omega) - H(\alpha\Omega)] \right). \quad (6)$$

### B. Upper Bounds

Theorem 1 shows that in many cases the sampling rate distortion function of a random source is equal to that of the arbitrary source. However, if  $\theta(\Omega, F)$  is less than the right hand side of (6), then the lower bound in Theorem 1 is less than the sparsity rate  $\Omega$  and there exists a gap with the upper bound given by the arbitrary setting (Proposition 1). In this section, we investigate improved (i.e. lower) upper bounds for these settings.

One way to upper bound  $\rho(\alpha)$  is to directly analyze the estimator that minimizes the error probability  $P_e^{(n)}$  given in (2). Although non-asymptotic properties of optimal estimation in the Gaussian setting have been studied (see for example [13]), analysis in the asymptotic setting appears to be challenging.

In this paper, we instead derive upper bounds for a computationally simple, and potentially suboptimal, estimator described below.

**Definition 6.** Suppose that the distribution of a random variable  $X$  is given by

$$X \sim \begin{cases} W, & \text{if } Z = 0 \\ W + \sqrt{\rho}U, & \text{if } Z = 1 \end{cases}$$

where  $U \sim F$ ,  $W \sim \mathcal{N}(0, \Omega \mathbb{E}[U^2])$ , and  $Z \sim \text{Bernoulli}(\Omega)$  are independent. For any subset  $T \subseteq \mathbb{R}$  let  $\hat{Z}_T(x) = \mathbf{1}(x \in T)$  and define the error probability

$$\epsilon(\rho, \Omega, F) = \inf_{T \subseteq \mathbb{R}} \Pr\{\hat{Z}_T(X) \neq Z\}. \quad (7)$$

**Definition 7.** For a random source  $\mathcal{X}(\Omega, F)$ , the *Thresholding* (TH) estimator  $\hat{\mathbf{s}}_{\text{TH}}(\mathbf{y})$  is given by

$$\hat{\mathbf{s}}_{\text{TH}}(\mathbf{y}) = \{i : \hat{\mathbf{u}}_i \in T^*\} \quad (8)$$

where  $\hat{\mathbf{u}} = B^T A^T \mathbf{y} \in \mathbb{R}^n$  and  $T^* \subseteq \mathbb{R}$  minimizes the right hand side of (7) with  $\rho = m/n$ .

The thresholding estimate corresponds to a separate hypothesis test for each element of  $\mathbf{x}$  and its complexity is linear in the vector length  $n$ .

**Proposition 3.** *Suppose that for each integer  $n$ , the elements of the sampling matrix  $\mathbf{A}$  are i.i.d.  $\sim \mathcal{N}(0, 1/n)$ . Then, for any random source  $\mathcal{X}(\Omega, F)$  and sampling rate  $\rho$ ,*

$$d(\hat{\mathbf{s}}_{\text{TH}}, \mathbf{S}) \rightarrow \epsilon(\rho, \Omega, F) \text{ in probability as } n \rightarrow \infty$$

where  $\epsilon(\rho, \Omega, F)$  is given by (7).

*Proof Sketch:* The key step, which is proved in [12], is to show that the empirical distributions of  $\{\hat{U}_i, i \in \mathbf{S}\}$  and  $\{\hat{U}_i, i \notin \mathbf{S}\}$  converge to the distribution of the random variable  $X$  described in Definition 6 conditioned on the events  $Z = 1$  and  $Z = 0$  respectively. ■

Combining Propositions 1 and 3 gives the following result which is complementary to Theorem 1.

**Theorem 2 (Upper Bound).** *A sampling rate distortion pair  $(\rho, \alpha)$  is achievable for the random source  $\mathcal{X}(\Omega, F)$  if  $\rho > \Omega$  or  $\alpha\Omega > \epsilon(\rho, \Omega, F)$  where  $\epsilon(\rho, \Omega, F)$  is given by (7).*

### C. A Gaussian Example

This section illustrates the bounds in Theorems 1 and 2 for a random source  $\mathcal{X}(\Omega, F)$  where  $F$  is a Gaussian distribution with mean  $\mu$  and variance  $1 - \mu^2$ . In Figure 2, the normalized sampling rate distortion function  $\rho(\alpha)/\Omega$  is plotted as a function of the mean  $\mu$  for  $\alpha = 0.3$ . It is shown that if  $\mu \leq \mu^* \approx 0.83$  then the number of samples needed is no different than for the arbitrary source  $\mathcal{X}(\Omega)$  signals. However, if  $\mu > \mu^*$ , there exists a gap between the bounds.

In Figure 3, the same bounds are shown for the relatively large distortion  $\alpha = 0.95$ . In this case, the upper bound from Theorem 2 is less than the rate needed for the arbitrary source, which verifies that, in some cases, there is a reduction in the number of samples that are needed. We note that the special case  $\mu = 1$  corresponds to a discrete distribution, and thus  $\rho(\alpha) = 0$  by Proposition 2.

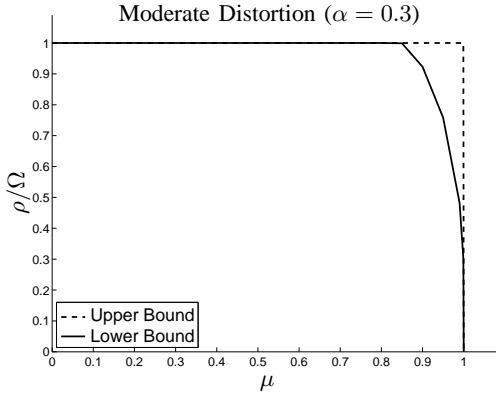


Fig. 2. Bounds on the normalized sampling rate  $\rho/\Omega$  needed to achieve distortion  $\alpha = 0.3$  as a function of the distribution mean  $\mu$  when the sparsity rate is  $\Omega = 0.35$  and the nonzero signal elements are i.i.d.  $\mathcal{N}(\mu, 1 - \mu^2)$ .

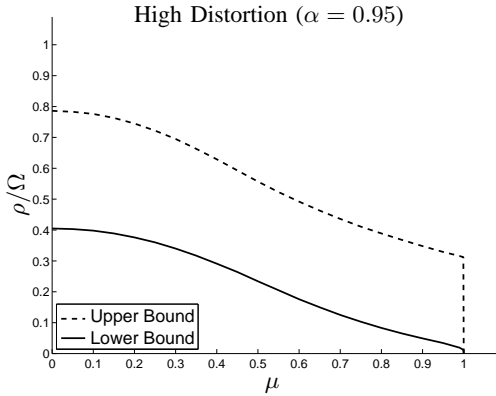


Fig. 3. Bounds on the normalized sampling rate  $\rho/\Omega$  needed to achieve distortion  $\alpha = 0.95$  as a function of the distribution mean  $\mu$  when the sparsity rate is  $\Omega = 0.35$  and the nonzero signal elements are i.i.d.  $\mathcal{N}(\mu, 1 - \mu^2)$ .

## V. PROOF OF THEOREM 1

Throughout this proof we use the notation  $A$  interchangeably to denote either a particular  $m_n \times n$  matrix  $A_n$  or a sequence of such matrices  $\{A_n\}$ . We begin with the following lemma which shows that the sampling rate distortion function can be lower bounded by considering an arbitrary sequence  $A$ .

**Lemma 1.** *Let  $A$  denote any sequence of full rank  $[\rho \cdot n] \times n$  sampling matrices. Then, for any distortion  $\alpha$ , the sampling rate distortion pair  $(\rho, \alpha)$  is not achievable for the random source  $\mathcal{X}(\Omega, F)$  if*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} I(\mathbf{AX}; \mathbf{S}|\mathbf{B}) < H(\Omega) - H(\alpha\Omega). \quad (9)$$

*Proof Sketch:* The lower bound for a given sequence  $A$  follows from Fano's inequality (see e.g. [11]). The fact that the bound for one matrix  $A$  applies to any other matrix  $A'$  (of equal rank) follows from that fact that there exists an invertible matrix  $D$  (based on the singular value decomposition) such that  $D\mathbf{AX}$  is equal in distribution to  $A'\mathbf{X}$ . ■

Next, we upper bound the left hand side of (9). Expanding the mutual information for a given problems size  $n$  gives

$$I(\mathbf{AX}; \mathbf{S}|\mathbf{B}) = h(\mathbf{AX}|\mathbf{B}) - h(\mathbf{AX}|\mathbf{S}, \mathbf{B}).$$

The entropy  $h(\mathbf{AX}|\mathbf{B})$  is upper bounded by the entropy of a Gaussian vector with the same covariance as  $\mathbf{AX}$ , and thus

$$h(\mathbf{AX}|\mathbf{B}) \leq \frac{m}{2} \log(2\pi e \sigma_x^2 |AA^T|^{\frac{1}{m}})$$

where  $\sigma_x^2 = \Omega\sigma_F^2 + \Omega(1-\Omega)\mu_F^2$  is the variance of each element of  $\mathbf{X}$ . Furthermore, the entropy  $h(\mathbf{AX}|\mathbf{S}, \mathbf{B})$  is lower bounded by the entropy power inequality [14] as

$$h(\mathbf{AX}|\mathbf{S}, \mathbf{B}) \geq \frac{m}{2} \mathbb{E} \log(2\pi e N(F) |\mathbf{AB}_\mathbf{S} \mathbf{B}_\mathbf{S}^T \mathbf{A}^T|^{\frac{1}{m}})$$

where  $N(F) = (2\pi e)^{-1} \exp(2h(F))$  is the entropy power of each nonzero element of  $\mathbf{U}$ . Combining these bounds gives

$$I(\mathbf{AX}; \mathbf{S}|\mathbf{B}) \leq \frac{m}{2} \mathbb{E} \log \left( \frac{1}{\theta(\Omega, F)} \cdot \frac{|AA^T|^{\frac{1}{m}}}{|\frac{1}{\Omega} \mathbf{AB}_\mathbf{S} \mathbf{B}_\mathbf{S}^T \mathbf{A}^T|^{\frac{1}{m}}} \right)$$

where we use the fact that  $\theta(F, \Omega) = \Omega N(F)/\sigma_x^2$ .

Without any loss of generality, we may assume that the spectral distribution of  $AA^T$  converges to a compactly supported probability measure  $\mu$  as  $n \rightarrow \infty$ . Then,  $|AA^T|^{\frac{1}{m}} \rightarrow G_\mu$  as  $n \rightarrow \infty$  where

$$G_\mu = \int_{\mathbb{R}} \log(x) d\mu(x).$$

The remaining problem, therefore, is to characterize the spectral distribution of the random matrix  $\mathbf{AB}_\mathbf{S} \mathbf{B}_\mathbf{S}^T \mathbf{A}^T$  as  $n$  becomes large. To this end, it is convenient to use results from free probability theory which is a theory for non-commutative probability theory developed by Voiculescu [15]. To begin, observe that the limiting spectral distribution of  $A^T A$  has a point mass  $\delta_0$  of weight  $1 - \rho$  at zero and is given by

$$\tilde{\mu} = (1 - \rho)\delta_0 + \rho\mu.$$

Observe also, that the limiting spectral distribution of  $\mathbf{B}_S^T \mathbf{B}_S$  is given by

$$\mu' = (1 - \rho/\Omega)\delta_0 + (\rho/\Omega)\delta_1$$

The basic idea from free probability is that the sequences  $A^T A$  and  $\mathbf{B}_S^T \mathbf{B}_S$  are *freely independent* and hence the spectral distribution of  $\mathbf{B}_S^T A^T A \mathbf{B}_S$  converges to a probability measure that can be described uniquely in terms of  $\mu$  and  $\mu'$ .

To characterize this measure, we use the following definition. The  $R$ -transform of a probability measure  $\mu$  is given by

$$R_\mu(z) = S_\mu^{-1}(-z) - \frac{1}{z}$$

where  $S_\mu^{-1}(z)$  denotes the inverse (with respect to the composition of functions) of the Stieltjes transform,

$$S_\mu(z) = \int_{\mathbb{R}} \frac{1}{x - z} d\mu(x).$$

The following result follows directly from Section 4.4 of Speicher's lecture on free probability [16].

**Lemma 2.** *If the limiting spectral distribution of  $A^T A$  is equal to  $\tilde{\mu}$ , then the limiting spectral distribution of the random matrix  $\frac{1}{\Omega} \mathbf{B}_S^T A^T A \mathbf{B}_S$  is equal to  $\tilde{\nu}$  almost surely where*

$$R_{\tilde{\nu}}(z) = R_{\tilde{\mu}}(\Omega z). \quad (10)$$

From Lemma 2 we conclude that spectral density of  $A \mathbf{B}_S \mathbf{B}_S^T A^T$  converges to  $\nu$  as  $n \rightarrow \infty$ , where

$$\nu = (\rho/\Omega - 1)\delta_0 + (\Omega/\rho)\tilde{\nu}.$$

If  $\nu$  is compactly supported then  $|\frac{1}{\Omega} A \mathbf{B}_S \mathbf{B}_S^T A^T|^{\frac{1}{m}} \rightarrow G_\nu$  almost surely as  $n \rightarrow \infty$ . Thus we conclude that

$$\limsup_{n \rightarrow \infty} I(\mathbf{A}\mathbf{X}; \mathbf{S}|\mathbf{B}) \leq \frac{m}{2} \log \left( \frac{1}{\theta(\Omega, F)} \cdot \frac{G_\mu}{G_\nu} \right) \quad (11)$$

almost surely for any compactly supported probability measures  $\mu, \nu$  that satisfy Equation (10).

Although the strongest bound corresponds to the minimization over  $\mu$ , such optimization appears to be difficult. Instead, we obtain a (potentially suboptimal) bound by setting  $\mu$  equal to the Marčenko-Pastur law [17] with parameter  $\rho$ , i.e.

$$d\mu(x) = \frac{\sqrt{(x-a)(b-x)}}{2\pi\rho x}$$

for all  $x \in [a, b]$  where  $a = (1 - \sqrt{\rho})^2$  and  $b = (1 + \sqrt{\rho})^2$ . Then, it can be shown that (10) is satisfied when  $\nu$  is equal to the Marčenko-Pastur law with parameter  $\rho/\Omega$ . Integrating with respect to these measures shows that

$$\begin{aligned} G_\mu &= e^{-1} \Delta(\rho) \\ G_\nu &= e^{-1} \Delta(\rho/\Omega) \end{aligned}$$

which completes the proof.

We remark that convergence of spectral density to the Marčenko-Pastur law corresponds to the setting where the elements of  $A$  are i.i.d. zero mean Gaussian. Interestingly, it is possible to use the rotational invariance of the Gaussian distribution to obtain the same bound given above, without

appealing to free probability. However, the approach taken above is more general and allows the calculation of the bound in terms of other limiting distributions.

## VI. DISCUSSION

Two insights from the field of compressed sensing are that any sparse vector can be sampled efficiently using linear projections, and that there exist random sampling matrices that good almost surely for any sparse basis. In this paper, we have investigated what happens if a probability measure is placed on the set of possible vectors and partial recovery is allowed by bounding the sampling rate distortion function  $\rho(\alpha)$ . In certain cases, we showed that the number of samples may be decreased. However, we also showed that in many cases, no reduction is possible, particularly if one requires universality with respect a sparse basis.

## ACKNOWLEDGMENT

This work was supported in part by ARO MURI No. W911NF-06-1-0076.

## REFERENCES

- [1] "Compressive sensing resources," <http://www.dsp.ece.rice.edu/cs>.
- [2] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," in *Proc. Allerton Conf. on Comm., Control, and Computing*, Monticello, IL, Sep. 2006.
- [3] —, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," in *Proc. IEEE Int. Symp. on Inform. Theory*, Nice, France, Jun. 2007.
- [4] G. Reeves, "Sparse signal sampling using noisy linear projections," Univ. of California, Berkeley, Dept. of Elec. Eng. and Cmp. Sci., Tech. Rep. UCB/EECS-2008-3, Jan. 2008.
- [5] S. Aeron, M. Zhao, and V. Saligrama, "On sensing capacity of sensor networks for the class of linear observation, fixed snr models," Jun 2007, arXiv:0704.3434v3 [cs.IT].
- [6] —, "Fundamental limits on sensing capacity for sensor networks and compressed sensing," Apr. 2008, arXiv:0804.3439v1 [cs.IT].
- [7] M. Akcakaya and V. Tarokh, "Shannon theoretic limits on noisy compressive sampling," Nov. 2007, arXiv:0711.0366v1 [cs.IT].
- [8] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," in *Proc. IEEE Int. Symp. on Inform. Theory*, Toronto, Canada, Jul. 2008.
- [9] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions on sparsity pattern recovery," May 2008, arXiv:0804.1839v1 [cs.IT].
- [10] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," in *Proc. IEEE Int. Symp. on Inform. Theory*, Toronto, Canada, Jul. 2008.
- [11] G. Reeves and M. Gastpar, "Approximate sparsity pattern recovery: Information-theoretic lower bounds," Preprint.
- [12] —, "Approximate sparsity pattern recovery: Information-theoretic upper bounds," Preprint.
- [13] —, "A note on optimal support recovery in a note on optimal support recovery in compressed sensing," in *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*, Monterey, CA, Nov. 2009.
- [14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [15] D. Voiculescu, "Asymptotically commuting finite rank unitary operators without commuting approximants," *Acta Sci. Math.*, vol. 45, pp. 429–432, 1983.
- [16] R. Speicher, "Free probability theory and random matrices," lectures at the summer school on 'Asymptotic Combinatorics with Application to Mathematical Physics', Jul. 2001.
- [17] V. Marčenko and L. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math. USSR-Sbornik*, vol. 1, pp. 457–483, 1967.